机器智能pk人类智能：大数据能够战胜人的逻辑推理吗？ | 文化纵横

原创 2016-03-09 吴军 文化纵横



吴军 / 谷歌公司
投稿微信：whzh_2014@126.com

> 大数据的这三个特点导致了机器智能和人具有完全不同的特点，它不是通过逻辑推理归纳演绎得出结论，而是利用大数据的完备性和多维度特点，直接找到答案。而大数据的完备性有可能让机器比人更能够掌控全局，或者说帮助决策者更好地掌握全局。

在过去的50多年里，人类GDP增长的根本动力是摩尔定律，即每过18个月，集成电路的性能（以集成电路芯片中的晶体管数量来衡量）就翻一番，或者说同样性能的集成电路每18个月价格下降一半。图1展示出不同年代集成电路芯片里晶体管的数量（纵轴，注意它是指数坐标）。在此之前，人类还没有一项技术能够在长达半个多世纪的时间里以指数增长的速度进步。集成电路的发展不仅开启了整个IT行业的技术革命，而且带来了全球的自动化和信息化，因此，这是在过去的半个世纪多里拉动世界经济增长的根本动力（虽然在中国还有房地产和基础设施建设，但是在世界范围内，这个市场是在萎缩的）。

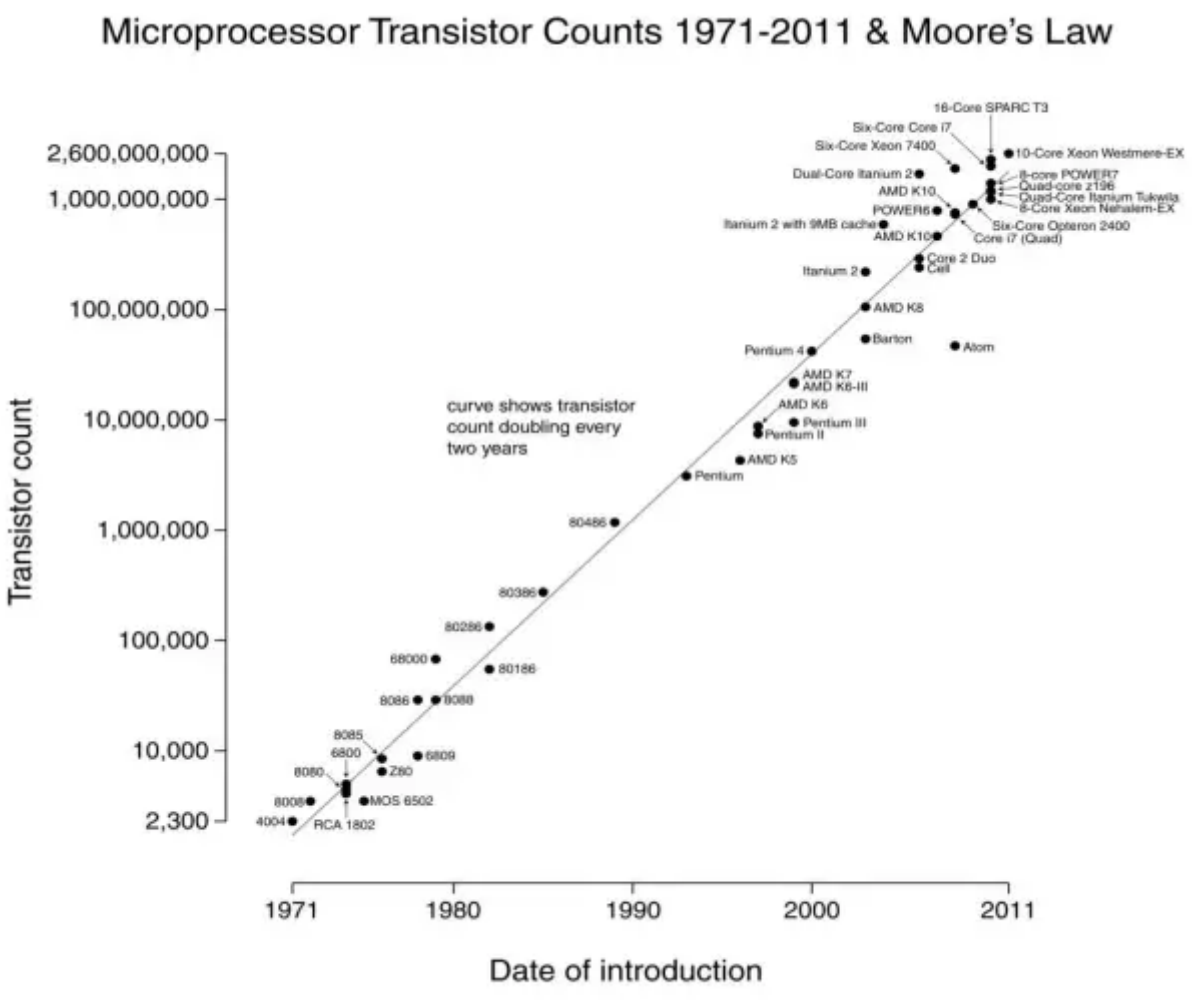


图1. 摩尔定律（横坐标是时间轴，纵坐标是集成电路中的晶体管数量）图中的点是不同时期具有代表性的处理器

摩尔定律带来的另外两个结果，就是互联网的兴起以及产业的数字化，而这两个结果合在一起，又产生了一个过去我们不太关注的结果，那是各种数据量的急剧增长。当数据量增加到一定程度，量变就有可能成为质变，因此今天大数据成为了一个非常热门的话题。不过，目前在行业里和社会上对大数据炒作居多，对于它的理解依然停留在比较浅的层面。这表现在：

1. 在概念上将大数据和大量数据相混淆，认为大数据就是数据量大，没有看到多维度和完备性的本质。
2. 在应用上（商业上）仅仅看到了利用统计规律提升业务，没有看到它和摩尔定律的结合必将导致机器智能社会的到来，从而彻底改变商业模式和产业结构，以及人们的生活和工作方式。
3. 明显低估了大数据和机器智能将给我们人类社会带来前所未有的机遇和冲击。

我们就从这三个方面来看看大数据、机器智能和它们对未来社会的影响。

可怕的大数据

要说清楚大数据的本质和作用，先要讲讲数据的作用以及它和机器智能的关系。

1. 数据驱动的方法导致机器智能的出现

从1946年计算机诞生以来，人类一直在思考这样几个问题：机器是否能有类似于人一样的智能？如果有，会在什么时候，以什么样的方式出现？如果出现了，它将对人类的生活产生什么影响？在过去的半个多世纪里，虽然人类一直在为此努力，但是到目前为止，这件事情似乎并没有发生，其中原因主要有两点。首先，人类在机器智能这个问题上一度过于乐观并且走了20年的弯路。人们最初的想法是让计算机来仿造智能，并且提出了人工智能的概念。学术界后来把这样的方法论称作“鸟飞派”，意思是说看看鸟是怎样飞的，就能模仿鸟造出飞机，而不需要了解空气动力学。（事实上我们知道，怀特兄弟发明飞机靠的是空气动力学而不是仿生学。）直到上个世纪70年代，人类才找到了适合计算机发展智能的道路，即采用数据驱动和大强度计算。其次，三十年前计算机的功能还不够强大，虽然当时人们以为它已经很快了，容量很大了，但是只有今天的十亿分之一左右，对解决智能问题来讲是远远不够的。而今天，人类正走在机器智能可能超越人类智能的转折点上。

讲到机器智能（而不是人工智能），首先要搞清楚什么是机器智能。1950年，计算机科学的先驱图灵博士给出了一个衡量机器是否有智能的测试方法：让一台机器和一个人坐在幕后，与一个人（测试者）展开对话（比如回答人的问题），当测试者无法被辨别和他讲话的是另一个人还是一台机器时，就可以认为这台机器具有和人等同的智能。这种方法被称为图灵测试 (Turing, 1959)。计算机科学家们认为，如果计算机实现了下面几件事情中的一件，就可以认为它有图灵所说的那种智能：

1. 语音识别，
2. 机器翻译，
3. 自动回答问题。

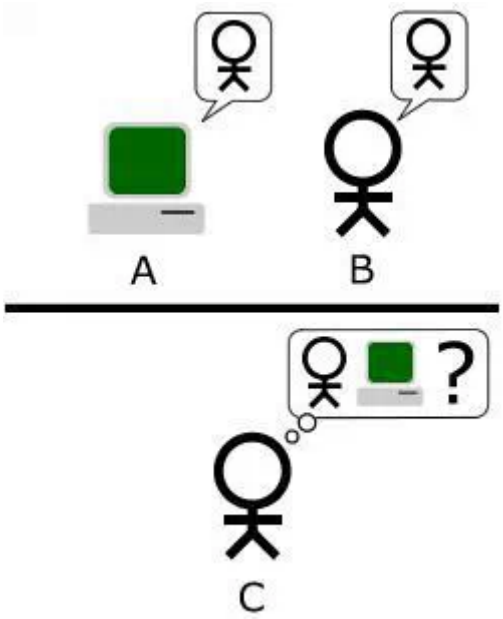


图 2 图灵测试

但是，从1950年代到60年代，机器智能按照传统人工智能的路子走得非常不顺利，几乎没有拿得出手的像样成果。而与此同时，计算机科学的其他分支都发展得非常迅速。因此，美国计算机学

界开始反思是否机器智能走错了路？1968年，著名计算机科学家明斯基在Semantic Information Process一书 (Minsky, 1968)中分析了所谓人工智能的局限性，他引用了Bar - Hillel使用过的一个非常简单的例子：

The pen was in the box.

这句话很好理解，如果让计算机理解它，做一个简单的语法分析即可。但是另一句话语法相同的话：

The box was in the pen.

就让人颇为费解了。原来，在英语中，Pen还有另外一个不太常用的意思 - - 小孩玩耍的围栏。在这里，理解成这个意思整个句子就通顺了。但是，如果用同样的语法分析，这两句话会得到相同的语法分析树，而仅仅根据这两句话本身，是无法判定pen在哪一句话中应该作为围栏，哪一句话应该是钢笔的意思。事实上，人对这两句话的理解并非来源于语法分析和语意本身，而来自于他们的常识，或者说关于世界的知识（World Knowledge），这个问题是传统的人工智能方法解决不了的。因此，明斯基给出了他的结论，“目前”（指当时）的方法无法让计算机真正有类似人的智能。由于明斯基在计算机科学界崇高的声望，他的这篇论文导致了美国政府削减了几乎全部人工智能研究的经费。在机器智能的发展史上，贾里尼克是一个划时代的人物。1972年，当时还是康奈尔大学教授的贾里尼克来到IBM沃森实验室进行学术休假，并且担任起IBM研制智能计算机的工作。贾里尼克于是挑选了一个他认为最有可能突破的课题，即语音识别。

贾里尼克从来不是真正的计算机科学家，而他的专长是信息论和通信，因此他看待语音识别问题完全不同于人工智能的专家们 - - 在他看来这是一个通信问题。人的大脑是一个信息源，从思考到合适的语句，再通过发音说出来，是一个编码的过程，经过媒介（空气或者电话线）传播到听众耳朵里，是经过了一个长长的信道的信息传播问题，最后听话人把它听懂，是一个解码的过程。既然是一个典型的通信问题，就可以用解决通信问题的方法来解决，为此贾里尼克用两个马尔可夫模型分别描述信源和信道。当然，为了训练和使用这两个马尔可夫模型，就需要使用大量的数据。采用马尔可夫模型，IBM 将当时的语音识别率从70%左右提高到90%以上，同时语音识别的规模从几百词上升到两万多词 (Jelinek, 1976)，这样，语音识别就能够从实验室走向实际应用。 贾里尼克和他的同事在无意中开创了一种采用统计的方法解决智能问题的途径，因为这种方法需要使用大量的数据，因此它又被称为是数据驱动的方法。

贾里尼克的同事彼得·布朗在1980年代，将这种数据驱动的方法用于了机器翻译 (P.F. Brown, 1990)。由于缺乏数据，最初的翻译结果并不令人满意，虽然一些学者认可这种方法，但是其他学者，尤其是早期从事这项工作的学者认为，解决机器翻译这样智能的问题，光靠基于数据的统计是不够的。因此，当时SysTran等公司依然在组织大量的人力，写机器翻译使用的语法规则。

如果说在1980年代还看不清楚布朗的方法和传统的人工智能的方法哪一个更适合计算机解决机器智能问题的话，那么在1990年代以后，数据的优势就凸显出来了。从1990年代中期之后的10年里，语音识别的错误率减少了一半，而机器翻译的准确性提高了一倍，其中20%左右的贡献来自于方法的改进，而80%则来自于数据量的提升。当然，这背后的一个原因是，由于互联网的普及，可使用的数据量呈指数增长。

最能够说明数据对解决机器翻译等智能问题的帮助的，是2005年NIST对全世界各家机器翻译系统评测的结果。

这一年，之前没有做过机器翻译的Google，不仅一举夺得了各项评比的第一名，而且将其它单位的系统远远抛在了后面。比如在阿拉伯语到英语翻译的封闭集测试中，Google系统的BLEU评分为51.31%，领先第二名将近 5%，而提高这五个百分点在过去需要研究7—10年；在开放集的测试中，Google51.37%的得分比第二名领先了17%，可以说整整领先了一代人的水平。当然，大家能想到的原因是它请到了世界著名的机器翻译专家弗朗兹·奥科(Franz Och)，但是参加评测的南加州大学系统和德国亚琛工学院系统也是奥科写的姊妹系统。从奥科在Google开始工作到提交评比结果，中间其实只有半年多的时间，奥科在方法上没有做任何改进。Google系统和之前的两个系统唯一的不同之处在于，前者使用了后者近万倍的数据量。

下表是2005年NIST评比的结果。值得一提的是，SysTran公司的系统是唯一采用传统的语法规则进行机器翻译的。它和那些采用数据驱动的系统相比，差距之大已经不在一个时代了。

从阿拉伯语到英语的翻译 （ 封闭集 ）

Google 51.31%
南加州大学 46.57%
IBM沃森实验室 46.46%
马里兰大学 44.97%
约翰·霍普金斯大学 43.48%
……
SYSTRAN公司 10.79%

从中文到英语翻译 （ 开放集 ）

Google 51.37%
SAKHR公司 34.03%
美军ARL研究所 22.57%

表1 2005年NIST对全世界多种机器翻译系统进行评比的结果

到了2000年之后，虽然还有一些旧式的学者死守着传统人工智能的方法不放，但是无论是学术界还是工业界，机器智能的主流方法是基于统计或者说数据驱动的方法。与此同时，另外两个相关的研究领域，机器学习和数据挖掘也开始热门起来。

2012 - 2014年，笔者曾经负责Google的机器问答项目，并且通过使用大数据，解决了30%左右的问题，这远远超过了学术界迄今为止同类研究的水平。究其原因，除了Google在自然语言处理等基础算法上做到了世界领先之外，更重要的是，Google将这个过去认为是存粹自然语言理解的问题变成了一个大数据的问题。首先，Google发现对于用户在互联网上问的各种复杂问题，有70-80%左右的问题可以在前十条自然搜索结果（去掉广告、图片和视频等结果）中找到答案，而只有20%左右的复杂问题，答案存在于搜索结果的摘要里。因此，Google将机器自动问答这样一

个难题转换成了在大数据中寻找答案的摘要问题。当然，这里面有三个前提，首先答案需要存在，这就是我们前面讲到的大数据的完备性；其次，计算能力需要足够，Google回答这样一个问题的时间小于10毫秒，但是需要上万台服务器同时工作；最后，就是要用到非常多的自然语言处理算法，包括对全部的搜索内容要进行语法分析和语义分析，要能够从文字的片段合成符合语法而且读起来通顺的自然语言等等。其中第一个前提是只有Google等少数大公司具备，而学术界不具备，因此这就决定了是Google而非学术界最早解决图灵留下的这个难题。

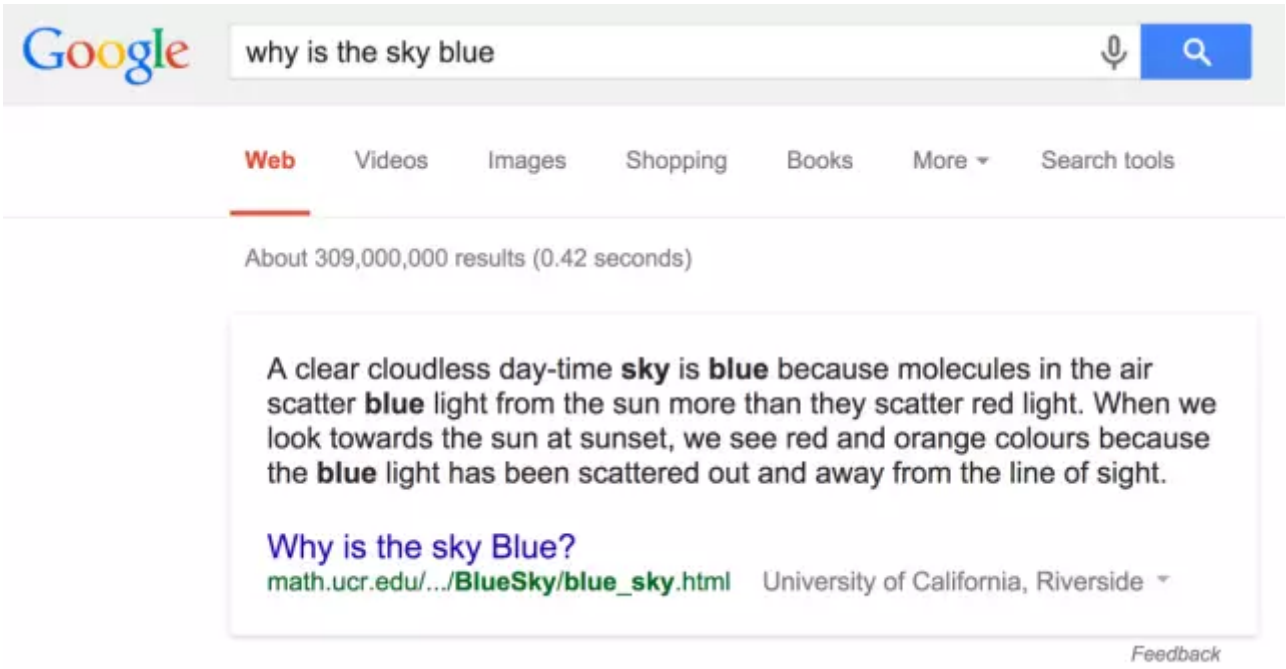


图 3 Google自动问答（问题为“天为什么是蓝色的？”，问题下面是计算机产生的答案）

由此可见，我们对数据重要性的认识不应该停留在统计、改进产品和销售，或者提供决策的支持上，而应该看到它（和摩尔定律、数学模型一起）导致了机器智能的产生。而机器一旦产生了和人类类似的智能，就将对人类社会产生重大的影响了。

2.大数据（Big Data）的本质

机器智能离不开数据，那么大量的数据和现在大家所说的大数据是否是一回事呢？如果不是，它们之间又有什么联系和区别呢?

毫无疑问，大数据的数据量自然是非常大的，但是光是“量”大还不是我们所说的大数据。比如过去国家统计局的数据量也很大，但是不是真正意义上的大数据。这两者的差别我们可以从三个方面来看。

首先，大数据具有多维度性质，而不同维度之间有着天然的（而非人为的）联系。为了说明这一点，我们不妨看一个实际的例子。

2013年9月份，百度发布了一个颇有意思的统计结果《中国十大“吃货”省市排行榜》。百度没有做任何的民意调查和各地饮食习惯的研究，它只是从“百度知道”的7700万条和吃有关的问题里“挖掘”出来一些结论：

在关于“什么能吃吗？”的问题中，福建、浙江、广东、四川等地的网友最经常问的是“什么虫能吃吗”，江苏、上海、北京等地的网友最经常问“什么的皮能不能吃”，内蒙古、新疆、西

藏，网友则是最关心“蘑菇能吃吗”，而宁夏网友最关心的竟然是“螃蟹能吃吗”。宁夏的网页关心的事情一定让福建的网友大跌眼镜，反过来也是一样，他们会惊讶于有人居然要吃虫子。

百度做的这件小事其实就是大数据的一个典型应用。它有这样一些特点。首先，它的数据量非常“大”。第二，这些数据维度其实非常多，它们不仅涉及到食物的做法、吃法、成分、营养价值、价格、问题来源的地域和时间等等，而且里面包含了提问者的很多信息，互联网的IP地址，所用的计算机（或者手机）的型号，浏览器的种类等等。这些维度也不是明确地给出的（这一点和传统的数据库不一样），因此在外面人看来，这些原始的数据是“相当杂乱”，但是恰恰是这些看上去杂乱无章的数据将原来看似无关的维度(时间、地域、食品、做法，成分，人的身份和收入情况等)联系了起来。经过对这些信息的挖掘，加工和整理，就得到了有意义的统计规律。

当然，百度只公布了一点点大家感兴趣的结果。它完全可以从这些数据中得到更多有价值的统计结果。比如，它很容易得到不同年龄人、性别和文化背景（这些很容易挖掘出来）的饮食习惯，不同生活习惯的人(比如正常作息的、夜猫子们、经常出差的或者不爱运动的等等)的饮食习惯等等。如果百度的数据收集的时间跨度足够长，它还可以看出不同地区人饮食习惯的变化，尤其是在不同经济发展阶段饮食习惯的改变。而这些看似很简单的问题，比如饮食习惯的变化，没有百度知道的大数据，还真难得到。这就是大数据多维度的威力。

大数据的第二个特点在于它的完备性。为了说明这一点，让我们再来看一个真实的案例。从1932年开始，盖洛普一直在对美国总统选举进行预测，几十年来它也在不断地改进采样的方法，力求使得统计准确，但是在过去的几十年里，它对美国大选结果的预测可以讲是大局（全国）尚准确，但是细节（每一个州）常常出错。因为再好的采样方法，也有考虑不周全之处。

但是到了2012年总统选举时，这种“永远预测不准”的情况得到了改变。一位名不见经传的统计学家Nate Silver通过对互联网网上能够取得的各种大量的数据（包括社交网络上用户发表的信息、新闻信息和其它网络信息），进行大数据分析，准确地预测了全部50个州的选举结果，而在历史上，盖洛普从来没有做的这一点。当然，有人可能会问，这个结果是否是蒙的？这个可能性或许存在，但是只有一千万亿分之一，因此可以认为这是大数据分析的结果。在这个例子中，Silver并没有什么好的采样方法，只是收集的数据很完备。大数据的完备性，不仅有用，甚至有点可怕。



图4. Nate Silver对2012年美国大选的预测（左）和实际结果（右）的对比（红色的地方表示共和党获胜的州，蓝色的表示民主党获胜的州）

数据的完备性的作用远比准确预测一个总统选举大得多，Google无人驾驶汽车便是一个很好的例子。首先，无人驾驶汽车可以算是一个机器人，这点应该没有疑问，因为它能像人一样对各种随机突发性事件快速地做出判断。在这个领域Google只花了六年时间就做到了全世界学术界几十年没有做到的事情。在2004年，经济学家们还认为司机是计算机难以取代人的几个行业之一。当然，他们不是凭空得出这个结论的，除了分析了技术上和心理上的难度外，还参考了当年DARPA组织的自动驾驶汽车拉力赛的结果－－当时排名第一的汽车花了几小时才开出8英里，然后就抛锚了。但是，仅仅过了6年后，2010年Google的自动驾驶汽车不仅研制出来了，而且已经在高速公路和繁华的市区行驶了14万英里，没有出一次事故。

为什么Google能在不到六年的时间里做到这一点呢？最根本的原因是它的思维方式和以往的科学家们都不同－－它把这个机器人的问题变成了一个大数据的问题。首先，自动驾驶汽车项目是Google街景项目的延伸，Google自动驾驶汽车只能去它“扫过街”的地方，而在行驶到这些地方时，它对周围的环境是非常了解的，这就是大数据完备性的威力。而过去那些研究所里研制的自动驾驶汽车，每到一处都要临时地识别目标，这是人思维的方式。其次，Google的自动驾驶汽车上面装了十几个传感器，每秒钟几十次的各种扫描，这不仅超过了人所谓的“眼观六路、耳听八方”，而且积攒下来的大量的数据，对各地的路况，以及不同交通状况下车辆行驶的模式有准确的了解，计算机学习这些“经验”的速度则远远比人快得多，这是大数据多维度的优势。这两点是过去学术界所不具备的条件，依靠它们，Google才能在非常短的时间里实现汽车的自动驾驶。

大数据的第三个特征在它的英文提法“Big Data”这个词当中体现的很清楚。请注意，这里使用的是Big Data，而不是Large Data。Big和Large这两个单词有什么区别呢，Big更主要是强调抽象意义上的大，而Large是强调数量（或者尺寸）大，比如大桌子Large Table。Big Data的提法，不仅表示大的数据量，更重要地是强调思维方式的不同。这种以数据为主的新的做法，在某种程度上颠覆了我们长期以来在科学和工程上的方法论。在过去，我们强调做一件事情的因果关系，通过前提和假设，推导出结果。但是在大数据时代，由于数据的完备性，我们常常是先知道结论，再找原因（甚至不去找原因），那么我们是否愿意去接受这样的工作方式。事实上，在一些具有大数据的IT公司里，包括Google，阿里巴巴等，今天已经是按照这种思维方式做事情了。Google的产品比竞争对手稍微好一点，主要不是靠技术，而是靠它的数据比对手更完备，同时它愿意用数据来解决问题。阿里巴巴的小额贷款能做起来（而其它商业银行做不到），其实就是对大数据思维的一种诠释。这是一种我们以前完全没见过的新的思维方式，一种新的方法论。

大数据的这三个特点导致了机器智能和人具有完全不同的特点，它不是通过逻辑推理归纳演绎得出结论，而是利用大数据的完备性和多维度特点，直接找到答案。而大数据的完备性有可能让机器比人更能够掌控全局，或者说帮助决策者更好地掌握全局。

大数据不仅仅是数据量大，而在于它的天然多维度特点和它的完备性。数据驱动的方法结合呈指数增长的计算机性能导致了机器智能的产生，并且在今天这个时间点上可以比肩人类的智能，这才是大数据重要的根本原因。机器智能和人的智能是不同的，它不是依靠人严密的逻辑推理得到

问题的答案，而是通过大数据的完备性直接找到答案，或者根据大数据多维度的特点找到以前我们无法发现的规律性。这将改变我们的思维方式，也就是所谓的采用“大数据思维”。

大数据和机器智能的井喷式爆发

大数据这个概念在今天这个时间点被提出来，其根本原因是因为摩尔定律导致互联网的发展，进而使得各种数据量的急剧增长（图5，数据增长的速度快于互联网本身增长的速度）。因此，大数据的第一个来源是互联网，包括移动互联网。大数据的产生还有第二个原因，就是传感器技术的突破 - - 今天各种手机，各种可穿戴式设备都有非常精准的传感器。而一些传感器价格非常便宜（比如RFID，每个售价仅四美分），使得每一个物品都可以装上一个。这些传感器产生了大量的数据。

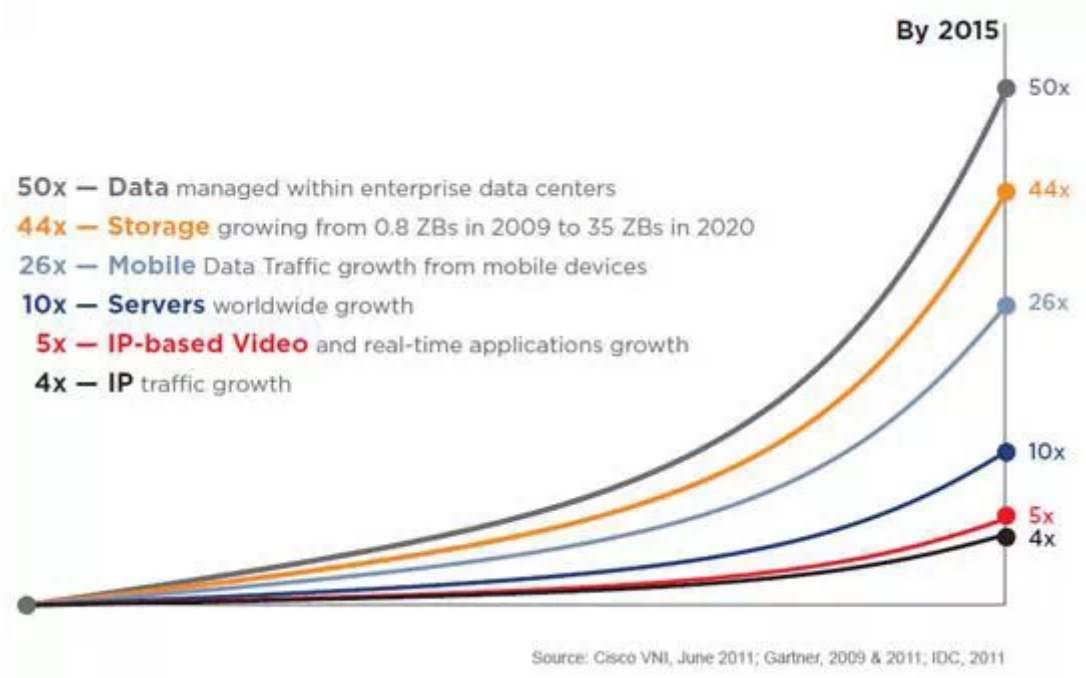


图5. 数据的增长速度（最上端的黑线）快于互联网的增长速度（数据来源：美国Gartner公司）

那么大数据能有多么完备？未来的机器能有多么“聪明”？我们看两个极端的例子。

1.精确到每一个细节

我们来构造两个场景，一个是关于物品，一个是关于人。

我们假定在未来的社会里，每一件物品上都贴有一个RFID（指甲盖大小的不干胶），那么它每经过一道（装有RFID阅读器）门，都可以记录下来，这样我们可以追踪它从出厂一直到被消费掉（或者最终销毁掉）的每一步。将来顾客在超市买东西时，他不再需要在收银台前排队，然后一件件扫描商品算账，而只需要把购物车推出装有RFID阅读器的大门，而那个阅读器将读出他购物车里每一件商品，并且算出价钱。不仅如此，他还能知道商品的来源，这样假货就难有藏身之处。对于厂家来讲，它的意义就更大了，大数据可以能建立起厂家和终端用户的直接联系。以前，厂家和客户之间或多或少隔着一些经销商，因此只能了解自己产品大致的销售情况（比如哪个地区卖了多少），而无法了解细节（比如哪个收入阶层在消费，各个流通环节加价多少等）。

但是在大数据时代，它不仅能够知道每一件商品买给了谁，甚至能知道中间每一个流通环节里的细节。因此，它可以完全根据市场供需进行生产，而且可以减少中间环节。对于政府的税收部门，如果可以备份每一笔交易的每一个细节，保证每一笔税收。当时，实现这一切需要多少IT投入，什么时候可以完成现在尚难估计。

接下来谈谈对人的精细化了解。从理论上讲，在大数据时代完全有条件了解每一个人24小时的全部行程。对于使用手机的人，这件事很容易做到。Google向智能手机和Chrome的用户提供一个被称为Google Now的个人资讯服务，它通过大数据（利用多维度的特点）将用户在日常生活中使用的很多（网络）服务打通，让使用者可以非常方便地管理每一天的生活。比如，它通过纪录使用者不同时间所在的位置，自动了解他的住址和上班的地点，每天上下班前，通知用户路上所需要的时间和比较好的行程路线，同时也能通知用户沿路可能发生的拥堵和交通事故。它还可以根据用户的通信（比如邮件），帮助用户自动地将每天的活动自动地加到日历中并且及时提醒下一个活动。事实上，用户在获得这种方便性的同时，将每天的活动全部交给了Google来管理。只要再做进一步的数据挖掘，拥有这样大数据的公司和机构（不一定需要是手机制造商和服务运营商），不仅可以知道一个人任何时刻的位置，而且可以知道他在做什么（比如在打电话，写邮件，开会或者工作）等等，甚至可以知道在什么时候他和什么人见了面（比如两个人在某个地方一起吃了一小时午饭，吃的是什么饭）。对于没有使用手机的人来讲，虽然不能获得如此详尽的信息，每天的活动也是有办法知道的。比如只要他身上任何一件物品是可以识别，或者他的一些生物特征（脸谱、指纹、声音等）可以识别，再经过大数据分析，也能比较详尽地了解这个人的活动。在过去，针对非常少量的人，如果不计成本的话，这件事情也能办到，但是不可能针对大范围的人，而在大数据和机器智能使得了解每一个人的生活变成了可能。

大数据和机器智能的发展，必将使得“机器”能够准确地了解社会的每一个细节。因此，具有最强大智能机器的不是哪一个具体的机器人，而是超级数据中心后面几十万，上百万的服务器集群。而掌控这个集群的人实际上在掌控这个社会发生的一切。

2.对医疗卫生的影响

近几十年来，虽然人类的寿命在不断地延长，但这在很大程度上是靠技术手段，而不是医疗本身水平的提高。事实上，提高医疗水平是一个非常漫长的过程，而过去研制新的药品和医疗手段亦是如此。医疗健保的费用不断增加，而且照此下去各国将是不堪负荷的，据估计到2020年，美国用于医疗健保的费用将达到GDP的20%左右。

那么出路在哪里？今天IT界和医学界领域都有一个共识，就是通过IT的进步，尤其是大数据和机器智能的进步，帮助解决人类健康的问题。当然，这个话题很大大，我们可以从三个侧面来看看在未来IT技术对生物医疗的帮助。

首先，药品的研制。攻克癌症是人类的一个梦想，但是迄今为止没有一种特效的抗癌药能够治愈癌症。过去医学界还试图研制这样的抗癌药，但是后来医学界认识到，由于癌细胞本身的基因会变异，因此并不存在这样一种万能药。基于这一点共识，医学界改变了治疗癌症的思路，那就是针对特定患者（不断变化的癌细胞），研制特定的药物，从理论上讲，只要研制的速度超过癌细胞变化的速度，癌症就可以治愈了。

按照传统的药品研发思路，科学家们应该先研究病理，找到解决方法（比如阻止具有某种基因的癌细胞蛋白质的合成），然后找到相应的药物，进行各种动物实验和临床试验。这是我们前面提到的强调因果关系的工作方法。但是，按照这个思路，为每一位癌症患者研制一种新药是很难办到的。且不说制药公司能否安排一个专门的团队为一个特定的患者服务，就算是能做到这一点，研制新药的成本也是患者无法负担的 - 平均一个人要十亿美元（基因泰克公司董事长李文森博士的估计）。事实上，不仅研制抗癌药成本高周期长，在美国，任何一种有效的处方药研制的时间和费用都非常高。过去大约需要十年时间，十亿美元，今天这个过程并没有缩短，而成本甚至上升到近百亿美元。

针对这种困境，科学家们想到了利用大数据来解决问题。在美国有大约5000多种处方药（远比一般人想象的少），过去每一种处方药都是用于当初针对的那些疾病的，比如治疗心脏病的药物就是用于心脏病的。但是，今天通过大数据统计研究发现，一款治疗心脏病的药物对于胃病的治疗效果明显。按照大数据的思维方式，我们应该先接受这个结论，再反过来找原因。基于这样的方法，找到治疗一种疾病的药品的组合，成本比以前研制新药成本要降低至少一个数量级，而时间可以缩短70-80%。根据著名生物系统专家、基因泰克公司董事长阿瑟·李文森博士的估计，采用大数据有望实现针对每一位癌症患者量身定制药物和治疗方法，而成本可以降到每个人五千美元。

其次，基因科技和医疗诊断。2012年Google科学比赛的第一名授予了一位来自威斯康星的高中生，她通过对760万个乳腺癌患者的样本数据的机器学习，设计了一种确定乳腺癌癌细胞位置的算法，来帮助医生对病人进行活检，其位置预测的准确率高达96%，超过目前专科医生的水平。可以讲，她的成功在很大程度上得益于大数据。这个例子只是众多通过IT技术来帮助疾病诊断的成功案例之一。一些类似的软件已经开始商用化。

大数据对医疗诊断的另一个主要的应用在于将人类的基因图谱和各种疾病联系起来，从而找到可能致病的基因并且设法修复。如果这件事情能够完成，那么不仅有希望治愈很多过去因为基因缺陷引起的绝症（比如癌症、帕金森综合症等），甚至有可能逆转人类的衰老过程。2013年，Google成立了它的医疗保健分公司Calico，并且聘请了李文森博士担任CEO，其第一期的投入已经高达10亿美元。据李文森博士介绍，采用传统的医学研究的方式，要想找到导致老年痴呆的基因并且找到治疗方法，在他有生之年（1950年出生）可能是看不到的，但是利用大数据，则有可能办到。

第三、医疗机器人。约翰·霍普金斯大学的罗素·泰勒教授是全世界最有权威的医疗机器人专家。根据他的专利制造的手术机器人达·芬奇已经成功地在全世界进行了150万例的手术，包括前列腺摘除，心脏瓣膜修复等。据泰勒教授介绍这种造价两百万美元的机器人采用了非常多的跨学科的技术，具体到IT领域，其核心技术包括图像处理和图像识别，3D图像的复原，统计模型等等。为了制造这个机器人，科学家们从医学院里收集了大量的数据，建立各种模型，从而让它拥有了一个非常见多识广的大脑。相比医生，手术机器人最大的好处在于它的稳定性 - - 即不会因为情绪而影响手术效果。从2000年这种机器人被FDA批准使用后，目前全球已经装备了3000多台，其中三分之二在美国。在未来，这一类的医疗机器人会越来越多地被使用。

第四、可穿戴式设备改变生活习惯。在大数据时代，可穿戴式设备将扮演很多角色，和以往互相不关联的电子设备不同，它有可能将人一天24小时都连到互联网上。使用者（和背后的大数据公司）可以通过可穿戴式设备了解到每他的生活习惯和健康状况。这可以为疾病诊断提供准确的数

据，试想一下，7天24小时监控的心跳和血压数据，一定比在医院一次测量的数据更加准确地反应了一个人的健康状况。每个人的健康数据将来还可以作为医疗保险收费的凭据，并且帮助人们养成一个良好的生活习惯。在美国，一些保险公司正在尝试给愿意使用可穿戴式设备，并且证明有良好生活习惯的人较低的保费。

可以预见，在不久的将来，IT技术尤其是大数据和机器智能技术在医疗保健上的应用会越来越多，并且将极大地改善人类的生活。

3.智能机器（机器人）

在2015年拉斯维加斯的消费电子产品展CES上，各厂家都用“所有东西皆智能（Smart Everything）”来吸引观众。当然，机器人（包括无人机）是展览会上的一个亮点。

无人机本身已经不是什么新鲜事，但是美国一家小公司SkyCatch则将它们的智能水平提到一个新的高度。这家公司让智能机的操作完全(比如换电池、换硬盘）由机器人来完成，用户只需要定义任务即可。苹果公司是SkyCatch的客户之一，目前租用无人机服务监控它的新总部施工情况，每天无人机要出动六、七次到工地上空拍摄出全部的工程进展录像和照片，能了解工地建设从第一天到最后一天全部的细节。

由于在地面为无人机服务的是是机器人，能派出的无人机的数量可以非常多，频率也可以非常高。如果我们把这件事想得远一点，无人机可以查出一个城市全部的违章建筑，任何地方的交通状况，对农业收成的估计、矿山和工地的勘察等等。如果把这种机器人能够做的事情再进一步扩展，变成为固定机翼的无人机装卸燃料，那么就构成了一直非常可怕的空军。

如果我们再把无人驾驶汽车看成一个种能在地面上行驶的机器人，那么它们经过改装，可以变成为反恐服务的巡逻车。当然，这些要以进一步的机器智能和大数据研究为基础。

当然，智能机器最关键的不在于可以移动的“四肢”，而在于它的大脑。在这超级“机器大脑”的控制下，智能机器可以完成很多人难以完成的事情，从在生产线上取代工人，到取代那些职业人士（金融、法律和医生）。这些在过去看似不可能的事情，今天正在发生。关于这一点，我们下面会详细论述。

大数据、机器智能和未来社会

机器智能无疑可以帮助改善人类的生活，包括延长人的寿命，但另一方面，我们在欢呼机器智能到来的同时，是否准备好了它对未来社会带来的冲击呢。技术革命的作用常常是正反两方面并存的。我们从对制造业、服务业和国家安全几个层面的影响来说明它。

2011年德国提出工业4.0的概念，即通过数字化和智能化来提升制造业的水平。其核心则是通过智能机器、大数据分析来帮助工人甚至取代工人，实现制造业的全面智能化。这在提高设计、制造和供应销售效率的同时，也会大大减少产业工人的数量。在中国，全球最大的OEM制造商富士康，一直在研制取代生产线上工人的工业机器人。富士康预计未来它将有上百万的机器人取代装

配线上的工人。这一方面使得工人们不再需要到生产线上去从事繁重而重复性的工作，另一方面则使得工厂里的工人数量将大幅度地减少。

当然，很多人会说，自从大机器出现后，工人的数量就在减少，但是劳动力会被分配到其他行业。但是，如同在2004年经济学家低估了机器可以取代驾驶员的可能性一样，今天我们可能在低估机器智能对未来社会的冲击。这一次由机器智能引发的技术革命，不仅仅是替代那些简单的劳动，而将在各个行业取代原有的从业人员，因为这将是人类历史上第一次，机器在智能方面超越人类。我们不妨看看机器智能对那些最需要专业技能的行业带来的冲击。

在美国，专科医生，比如放射科医生，是社会地位和收入最高的群体，也是需要专业知识最多，智力水平最高的群体 - - 他们需要在大学和医院学习和训练13年（平均）才能获得行医的执照。这样的工作，过去被认为是不可能被机器取代的。但是，今天智能的模式识别软件通过医学影像的识别和分析，可以比有经验的放射科医生更好地诊断病情，而这个成本，只是人工的百分之一。

律师，也被认为是最“高大上”的职业，但是他们的工作受到了自然语言处理软件的威胁。今天，打一场像苹果vs三星这样的官司，要分析和处理上百万份法律文件（因为美国是判例型法律），律师费高得惊人。但是，位于硅谷Palo Alto的 Blackstone Discovery公司发明了一种处理法律文件的自然语言处理软件，使得律师的效率可以提高500倍，而打官司的成本可以下降99%。这意味着未来将有相当多的律师可能失去工作。事实上这件事情在美国已经发生，新毕业的法学院学生找到正式工作的时间比以前长了很多。

面对势不可挡的机器智能大潮，人类在未来需要重新考虑工作和生活的方式了，尤其是劳动力的出路问题。从一百多年前开始的农业革命使得发达国家2-5% [注释：根据美国劳工部的统计，美国农业工人占不到劳动力人口的2%。] 的人提供了全部人口所需的食品，随着机器智能的发展，或许只需要5%以下的劳动力就能提供人类所需的所有工业品和大部分的服务工作。当然，会有一小部分人参与智能机器的研发和制造，但是这只会占到劳动力的很小一部分。那么，我们现在必须考虑未来劳动力的出路在哪里？这是一个在机器智能发展过程中无法回避的问题。

在历史上，第一次工业革命（1760-1820）一方面极大地解放了生产力，并且使得人类第一次生产的产品超过了消费的需求。当时在世界上真正受益的只有英国，起产业工人不过数百万而已。即便如此，它也感觉到市场不够用，以至于它需要进行全球性市场的拓展。亚当斯密的《国富论》诞生于那个年代有它的历史背景。在这本经典经济学著作中，亚当斯密阐述了英国开放市场的重要性，而其目的是换取他国同样地开放市场。这次工业革命进行了六十年，有两代人的时间消化它带来的冲击力，但是在很长时间里，社会依然无法承受。在19世纪上半叶，是英国劳资矛盾最突出的时期，狄更斯等人的小说里描写的工人们悲惨的景象，便是在那个时期，而马克思主义也恰恰是在那个时代诞生的。直到19世纪中，英国才形成一个相对稳定的中产阶级群体，而靠着对外贸易，英国在19世纪中期进入它的维多利亚荣光时代，其标志为1855年首届世博会。如果从1760年算起，英国整整花了一个世纪才消化工业革命带来的负面影响，这还是在全世界独此一国进入工业化时代的前提下。

第二次工业革命（1870-1914）也历时了近半个世纪，电的使用帮助美国和德国超越了英国，并且成为人类历史上财富增长最快的时期。在美国一方面造就了范德比尔特、卡内基和洛克菲勒等

商业巨子（在全世界有史以来最富有的75人中，有五分之一出现在第二次工业革命时期的美国），但是另一方面，这也是美国社会公平性最受破坏的事情。最后，经过老罗斯福、塔夫特和威尔逊三任总统反垄断的努力以求恢复社会的公平性、经过了一次大战从债务国变成债权国，直到20世纪20年代，才迎来了柯立芝繁荣。这前后也花了半个多世纪来适应这次产业革命。

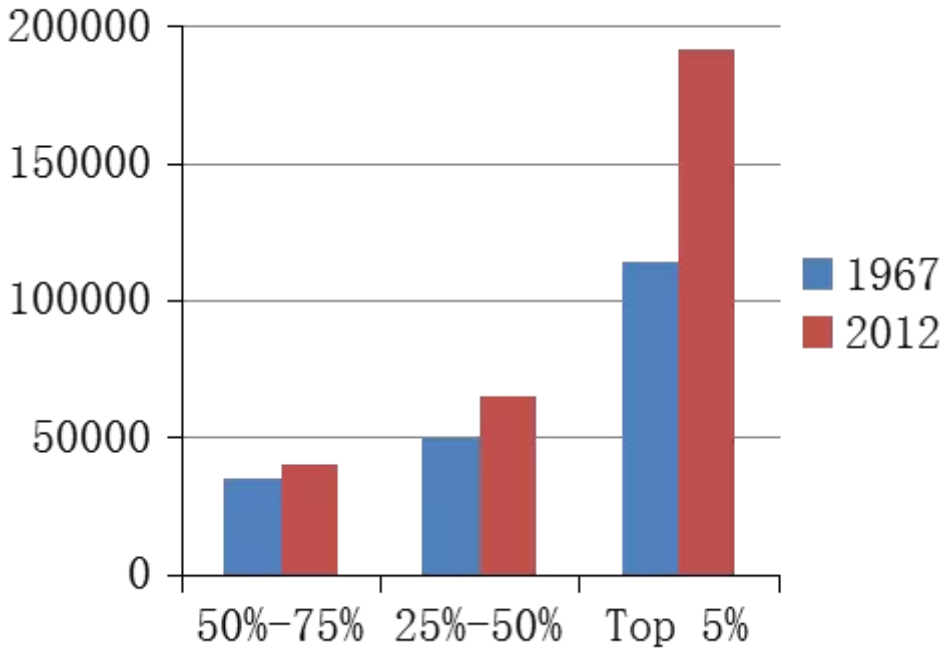


图3. 从1967到2012年美国家庭收入变化

在从上个世纪末开始的所谓第三次浪潮的信息革命中，美国无疑是一个领跑者和赢家，但是美国大部分民众在经济上和幸福指数上其实并没有什么提升（图3，美国家庭在扣除通货膨胀后收入的增长，最左边是收入排在50% - 75%的家庭，中间是排在25% - 50%的家庭，右边是最富有的5%的家庭）和（图4，美国人幸福指数的变化）。

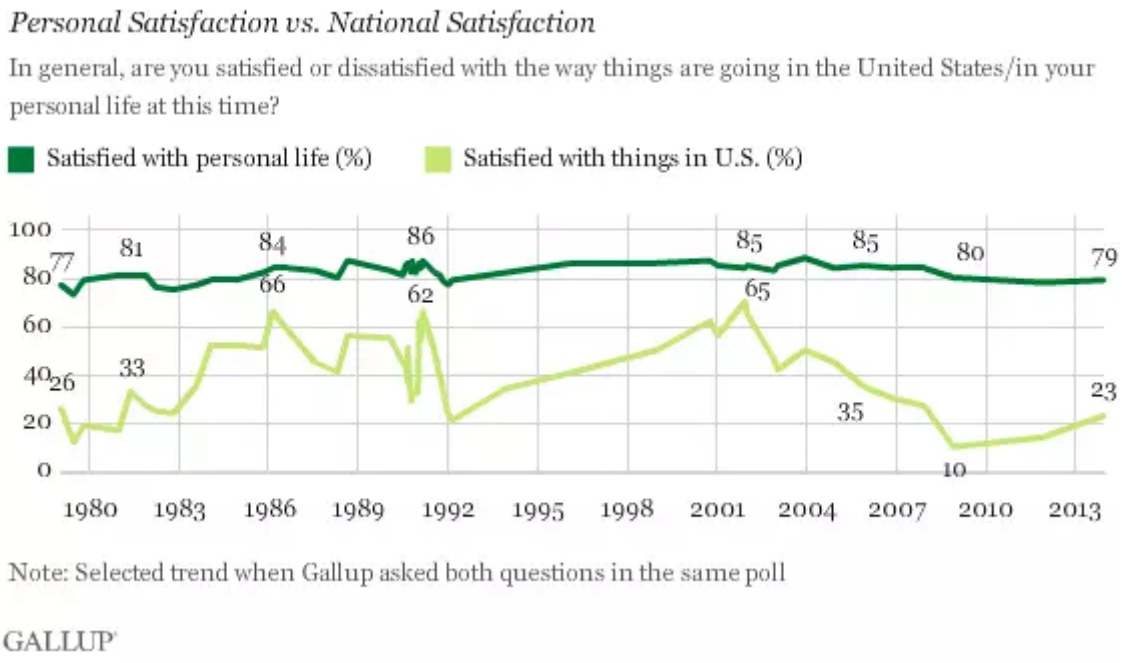


图4. 美国人幸福指数的变化（深色线是对个人生活的满意程度，浅色线是对物质生活的满意程度）

这一次将由机器智能带了的革命，对社会的冲击将是全方位的。社会结构可能会发生根本性的变化，这可能不是简单地把农业人口变成城市人口，把第一第二产业变成第三产业这么简单。第一次工业革命和第二次工业革命还有半个世纪左右的时间消化吸收这些变化，但是这一次由于机器智能和大数据带来的革命来得非常之快，涉及的领域非常之多，以至于我们未必会有很长的时间来应对，社会将如何适应这种变化，是决策者现在需要考虑的问题。

本文刊于《文化纵横》2015年4月号。原题为“大数据、机器智能和未来社会”图片来自网络。版权所有，欢迎个人分享，媒体转载请回复本公号获得许可。

点击左下角“阅读原文”，可便捷下单购买《文化纵横》2016年2月号纸刊。
亦可到淘宝网搜索店铺“文化纵横杂志社”购买各期纸刊：https://shop130801728.taobao.com



阅读原文